Frequentist Post-Data Inference

BU-1202-M April 1993

Constantinos Goutis
University College London

George Casella[1]
Cornell University

AMS 1980 subject classifications. Primary 62C05; Secondary 62A20.

Key words and phrases: Accuracy estimation, conditional confidence, conditional evaluation, confidence intervals, decision theory, hypothesis testing, loss estimation, relevant sets, sample space partitions.

[1] Research supported by National Science Foundation Grant No. DMS9100839 and National Security Agency Grant No. 90F-073.

Summary

The end result of an experiment is an inference, which is typically made after the data have been seen (a post-data inference). Classical frequency theory has evolved around pre-data inferences, those that can be made in the planning stages of an experiment, before data are collected. Such pre-data inferences are often not reasonable as post-data inferences, leaving a frequentist with no inference conditional on the observed data. We review the various methodologies that have been suggested for frequentist post-data inference, and show how recent results have given us a very reasonable methodology. We also discuss how the pre-data/post-data distinction fits in with, and subsumes, the Bayesian/frequentist distinction.

1. Introduction

Historically, statistical methodologies have evolved through two major schools of thought: Bayesian and frequentist. Bayesian statistical methods result in inferences that are conditional on the observed data, while frequentist methods result in inferences that are unconditional on the observed data. Stated in another way, Bayesian inferences are typically post-data inferences, that is, they are only made after the data have been seen. In contrast, frequentist inferences are pre-data inferences, as the classical frequentist probability structure only exists before the data have been seen. This is illustrated in the following example.

Example 1. Suppose that we observe $x_1, x_2, \ldots, x_n$, realised values of independent random variables $X_1, X_2, \ldots, X_n$ each assumed to have a normal distribution with mean μ and variance σ^2, that is, $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$. (Note the distinction that capital letters denote an unobserved random variable while lower case letters denote realised values). From $X_1, X_2, \ldots, X_n$ we can calculate $\bar{X}$ and S^2, the sample mean and variance and assert that the random interval

$$C_t(\bar{X}, S^2) = \left\{ \mu : \bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \right\} \tag{1.1}$$

will cover μ with probability $1-\alpha$ (as $t_{n-1,\alpha/2}$ is the appropriate Student's t cutoff point). This is a pre-data inference about the procedure and is valid before the data are seen.

Suppose we have $n = 5$ and $1-\alpha = 0.9$ (hence $t_{4,\alpha/2} = 2.132$). We now observe values $\bar{x} = 217.2$ and $s = 31.3$, and construct the interval

$$\begin{aligned} C_t(\bar{x}, s^2) &= \left\{ \mu : 217.2 - 2.132\,\frac{31.3}{\sqrt{5}} \le \mu \le 217.2 + 2.132\,\frac{31.3}{\sqrt{5}} \right\} \\ &= \{ \mu : 187.36 \le \mu \le 247.04 \}. \end{aligned} \tag{1.2}$$

The pre-data inference for the interval (1.1) does not apply to the interval (1.2). Any assertion about the coverage of the realised interval (1.2) is necessarily a post-data inference. ||

Although the Bayesian inference is typically a post-data inference, there may be cases when a Bayesian would be interested in a pre-data inference. This would occur in the planning stages of an experiment (before data are collected) and, in such a case, a Bayesian would perform a frequency-like pre-data inference (involving integrating over the sample space).

Example 1 (continued). To evaluate $C_t(\bar{X}, S)$ we consider the average performance using a density for $(\bar{X}, S)$. Let $f(u, v \mid \mu, \sigma)$ denote the sampling density. Then the frequentist would calculate the pre-data quantity

$$P(\mu \in C_t(\bar{X}, S) \mid \mu, \sigma) = \int_0^\infty \int_{C_t(u,v)} f(u, v \mid \mu, \sigma)\, du\, dv . \tag{1.3}$$

The Bayesian pre-data evaluation would be based on the marginal distribution of $(\bar{X}, S)$, given by

$$m_\pi(u, v) = \int_0^\infty \int_{-\infty}^\infty f(u, v \mid \mu, \sigma)\, \pi(\mu, \sigma)\, d\mu\, d\sigma \tag{1.4}$$

where $\pi(\mu, \sigma)$ is a chosen prior. The Bayesian pre-data inference is then based on the calculation

$$P(\mu \in C_t(\bar{X}, S)) = \int_0^\infty \int_{C_t(u,v)} m_\pi(u, v)\, du\, dv. \tag{1.5}$$

||

In contrast, the frequentist has no standard methodology for making a post-data inference. Whether or not the data have been seen, the standard frequentist inference is a pre-data one.

Example 2. In testing $H_0: \mu \le 0$ vs. $H_1: \mu > 0$ for the situation in Example 1, if $\phi(\cdot)$ denotes the probability of rejection as a function of the data, a frequentist procedure could be based on the rejection rule

$$\phi(\bar{x}, s) = \begin{cases} 1 & \text{if } \frac{\bar{x}}{s} > k \\ 0 & \text{if } \frac{\bar{x}}{s} \le k \end{cases}, \tag{1.6}$$

where k is chosen to give this procedure a prespecified Type I error. This is pre-data inference, and is the only inference available for the classical frequentist, even if the data have been seen. In contrast a Bayesian might calculate the posterior probability of H_0,

$$\begin{aligned} P(\mu \le 0 \mid \bar{X} = \bar{x}, S = s) &= \int_0^\infty \int_{-\infty}^0 \frac{f(\bar{x}, s \mid \mu, \sigma)\, \pi(\mu, \sigma)}{m_\pi(\bar{x}, s)}\, d\mu\, d\sigma \\ &= \int_0^\infty \int_{-\infty}^0 \pi(\mu, \sigma \mid \bar{x}, s)\, d\mu\, d\sigma , \end{aligned} \tag{1.7}$$

where $\pi(\mu, \sigma \mid \bar{x}, s)$ is the posterior distribution. ||

Experimenters tend to favour post-data inferences, as such inferences will often reflect how

strongly the data support a conclusion. This is the reason, perhaps, that a measure such as the p-value has gained widespread use. Interestingly, experimenters who would not consider a Bayesian analysis because of its "dependence on subjective probability" would report a p-value as a post-data measure of evidence. Of course, the p-value can be interpreted in a Bayesian way and, as such, does reflect a subjective belief. Such a belief may not be a plausible one, and many Bayesians argue that such "hidden" subjectivity is bad. However, it can alternately be argued that the p-value exists independent of subjective concerns. (These opinions we neither support nor condemn, but merely report.)

What this leads to is the necessity for frequentist post-data inference. More precisely, a methodology for constructing and assessing post-data accuracy estimates in a manner that is consistent with frequency theory. A number of such methodologies do exist, and in this paper we will review and explain them, and illustrate reasonable methods for performing such inference.

In the following sections we review the development of, and describe methodologies for performing frequentist post-data inference. We start, in Section 2, with the subject of conditional performance of frequentist procedures, which demonstrate the folly of using pre-data measures for post-data inference. Then, in Section 3, we discuss the methodology of Kiefer (1977), a first attempt to construct frequentist post-data measures. Kiefer's method, unfortunately, has not provided a workable solution, perhaps because of difficulties (and ambiguities) in applications. A more successful methodology, perhaps because of its simplicity, is estimation of indicator functions, detailed in Section 4. This methodology is mainly due to Berger (1985a), but had some seeds in Kiefer (1977) and Robinson (1979a). In Section 5, we discuss generalisations of this methodology, which might be called estimation of accuracy. Lastly, Section 6 contains a discussion that, we hope, serves to place this entire subject in perspective.

2. Conditional Evaluations

Classical frequentist evaluations are concerned with long-run behaviour and involve unconditional evaluations of a procedure. While such evaluations are important, they may possibly mask undesirable conditional behaviour. That is, even if a procedure is optimal unconditionally (averaged over all data), there may be particular data partitions on which its performance is suspect. Such investigations have illustrated the futility of using pre-data inferences in a post-data manner, and laid the foundations for constructing frequentist post-data measures.

2.1. The Need for Conditional Measures

The following two simple examples illustrate the fact that good pre-data frequentist measures are not necessarily good post-data measures.

Example 3. (Berger and Wolpert 1988). Let X be a p-variate normal random variable with mean θ and identity covariance matrix. The celebrated James-Stein estimator (James and Stein 1961)

$$\delta(x) = \left(1 - \frac{p-2}{\|x\|^2}\right)x \tag{2.1}$$

dominates x as an estimate of the mean θ for dimension larger than 3, using expected distance (mean squared error) as a criterion. However if $x = (0.01,\ 0.01,\ 0.01)$ then $\delta(x) = (-33.323,\ -33.323,\ -33.323)$. This seems unreasonable and we are almost certain that for this particular realisation of X, the maximum likelihood estimate, x, is superior to $\delta(x)$, in the sense of being closer to the true mean. Of course there is a tiny chance that

$$\|\delta(x) - \theta\|^2 < \|x - \theta\|^2 , \tag{2.2}$$

but it seems unlikely. Note that within the frequentist framework, we cannot talk about the probability of (2.2) since the realised distances are not random but fixed unknown numbers. ||

Example 4. Suppose that $X_1,\ X_2\ ,\ldots,\ X_n \sim U(\theta - 1/2,\ \theta + 1/2)$, independently. A $1-\alpha$ confidence interval for θ is given by

$$C_U(x) = \left(\frac{x_{(n)} + x_{(1)}}{2} - \frac{1 - \sqrt[n]{\alpha}}{2},\ \frac{x_{(n)} + x_{(1)}}{2} + \frac{1 - \sqrt[n]{\alpha}}{2} \right), \tag{2.3}$$

where $x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$ and $x_{(1)} = \min\{x_{1,2}, \ldots, x_n\}$. Suppose $1-\alpha = 0.9$, $n = 10$, $x_{(1)} = 0.01$

and $x_{(10)} = 0.99$, so the realised confidence interval is $C_U(\mathbf{x}) = (0.397,\ 0.603)$. However, for these values of $x_{(1)}$ and $x_{(10)}$ we are absolutely certain that $0.49 \le \theta \le 0.51$, hence the interval covers the parameter with certainty. So either the statement "we are 90% confident that $\theta \in C_U(X)$" is an understatement, or the realised interval $C_U(x)$ is too wide for its purpose. Had the sample been $X_{(1)} = 0.49$ and $X_{(10)} = 0.51$, the interval (2.3) would be exactly the same but we would be considerably less certain as to whether the true parameter is covered. ||

The problems in the above examples are obvious. The statements "$\delta(X)$ dominates X for all θ" and "$C_U(X)$ is a $1-\alpha$ confidence interval" are correct pre-data statements, in the sense that they are true if we use integration over the whole sample space. However, for selected subsets of the sample space they need not be correct. It can be proved (see the Appendix) that for Example 3

$$E_\theta\Big(\|\delta(X)-\theta\|^2 \,\Big|\, \|X\|^2 \le \epsilon \Big) > E_\theta\Big(\|X-\theta\|^2 \,\Big|\, \|X\|^2 \le \epsilon \Big) \tag{2.4}$$

for all θ and sufficiently small ϵ whereas

$$P_\theta\Big(\theta \in C_U(X) \,\Big|\, X_{(n)}-X_{(1)} \ge \epsilon_1\Big) > 1-\alpha \tag{2.5}$$

and

$$P_\theta\Big(\theta \in C_U(X) \,\Big|\, X_{(n)}-X_{(1)} \le \epsilon_2\Big) < 1-\alpha \tag{2.6}$$

for appropriate ϵ_1 and ϵ_2 and all parameters θ.

Thus, in both cases, the problems are easy to solve by slightly modifying the procedure, in the first case by taking the positive part James Stein estimator and, in the second, by conditioning on the ancillary $X_{(n)}-X_{(1)}$. Note that the expressions in (2.4) – (2.6) are, in some sense, post-data statements, although they involve integration over part of the sample space. We know something about the risk or the coverage probability if the event on which we condition is realised, which can be verified only after the data are collected. It is also interesting to note that one need not have the exact values of the data; the only information needed is in which element of a (coarse) partition of the sample space they belong.

2.2. Reference Sets

The fact that pre-data inferences may be substantially altered on different data partitions has been known for a long time. As with most ideas in statistics, the idea of conditional evaluations of an inference from a procedure can be traced back to Fisher. In particular, Fisher (1959) first noticed a problem with the Aspin-Welsh solution to the Behrens-Fisher problem, that the proposed test statistic did not retain its nominal level when conditioned on a *recognisable* subset (a specific set in the sample space). Fisher, of course, realised the importance of the concept of conditional inference but, other than a few examples, he left the future generations of statisticians with only a vague statement to formalise. Several researchers (Buehler 1959, Wallace 1959, Pierce 1973, Robinson 1975) examined conditional inference of confidence intervals, with Robinson (1979a,b) formalising and generalising the theory.

A version of the conditional inference question can be stated as follows: For a given confidence interval C(X) with a coverage probability $\gamma(\theta)$, or confidence coefficient $\gamma = \inf_{\theta} P\Big(\theta \in C(X)\Big)$, is there a subset $\mathcal{A}$ of the sample space such that

$$P_\theta\Big(\theta \in C(X) \,\Big|\, X \in \mathcal{A}\Big) - \gamma > \epsilon \tag{2.7}$$

or

$$P_\theta\Big(\theta \in C(X) \,\Big|\, X \in \mathcal{A}\Big) - \gamma < -\epsilon \tag{2.8}$$

for all parameter values θ and a positive, or at least non-negative ϵ? Such a set $\mathcal{A}$ is called a *relevant set* for the confidence interval C(X). The existence of such a set is a deficiency, for it is immediate to construct a better estimate of confidence [as in (3.1)].

Conditional inference properties and evaluations can be generalised by considering any measure of confidence $\gamma(x)$ (that is, not necessarily coverage probability), possibly dependent on x. By using betting scenarios, the left-hand side of (2.7) need not be conditional probabilities but weighted expectations of indicator functions.

In particular, if we were to assert confidence $\gamma(x)$ in the set C(x) (which we abbreviate as the confidence procedure $\langle C(x), \gamma(x)\rangle$), we would say that $\mathcal{A}$ is a relevant set for $\langle C(x), \gamma(x)\rangle$ if either

$$P_\theta\Big(\theta \in C(X) \,\Big|\, X \in \mathcal{A}\Big) - E_0\Big(\gamma(X) \,\Big|\, X \in \mathcal{A}\Big) \geq \epsilon \tag{2.9}$$

or

$$P_\theta\Big(\theta \in C(X) \,|\, X \in \mathcal{A}\Big) - E_0\Big(\gamma(X) \,|\, X \in \mathcal{A}\Big) \leq \epsilon \tag{2.10}$$

for some $\epsilon > 0$ and all θ. This can be further generalized to functions more general than indicator functions, however such a generalisation seems of lesser statistical interest. The interest of the question posed as above lies in that if (2.7) or (2.8) is true, we should be suspicious about quoting confidence equal to γ or $\gamma(x)$ if it happens that $x \in \mathcal{A}$. For several well known confidence procedures such sets exist, whereas for others it has been proved that one cannot find such sets (Buehler and Feddersen 1963, Brown 1967, Robinson 1975, Maatta and Casella 1987, Olshen 1973). For a review of this approach see Casella (1992).

An immediate question that emerges for the confidence set scenario is to characterize when there exists sets $\mathcal{A}$ satisfying (2.7) or (2.8). A partial answer was given by Pierce (1973) and Robinson (1979a). Roughly speaking, we can say that no set $\mathcal{A}$ will exist if we quote confidence $\gamma(x)$ equal to the posterior probability of C(X) with respect to some (possibly generalised) prior.

To see why this is so, suppose that we assign confidence $\gamma^\pi(x)$ to the set C(x) by calculating

$$\gamma^\pi(x) = \int_{C(x)} \pi(\theta \,|\, x)\, d\theta \,, \tag{2.11}$$

where $\pi(\theta \,|\, x) = f(x \,|\, \theta)\,\pi(\theta) \Big/ \int f(x \,|\, \theta)\,\pi(\theta)\, d\theta$, the posterior distribution that results from the prior $\pi(\theta)$. Now suppose that there exists a set $\mathcal{A}$ that satisfies the inequality in (2.9) for the confidence procedure $\langle C(x),\, \gamma^\pi(x)\rangle$. Expand the integrals and rewrite the inequality as

$$\int_{\mathcal{A}} \Big[I\Big(\theta \in C(x)\Big) - \gamma^\pi(x) \Big] f(x \,|\, \theta)\, dx \geq \epsilon \int_{\mathcal{A}} f(x \,|\, \theta)\, dx \,.$$

Now integrate both sides against $\pi(\theta)$ to get

$$\int_\Theta \int_{\mathcal{A}} \Big[I\Big(\theta \in C(x)\Big) - \gamma^\pi(x) \Big] f(x \,|\, \theta)\, dx\, \pi(\theta)\, d\theta \geq \epsilon \int_\Theta \int_{\mathcal{A}} f(x \,|\, \theta)\, \pi(\theta)\, dx\, d\theta \,. \tag{2.12}$$

On the left-hand side of (2.12), write $f(x \,|\, \theta)\,\pi(\theta) = \pi(\theta \,|\, x)\, m(x)$, where m(x) is the marginal distribution of X. Then interchange the order of integration and the inequality becomes

$$\int_{\mathcal{A}} \left\{ \int_\Theta \Big[I\Big(\theta \in C(x)\Big) - \gamma^\pi(x) \Big] \pi(x \,|\, \theta)\, d\theta \right\} dx \geq \epsilon \int_\Theta \int_{\mathcal{A}} f(x \,|\, \theta)\, \pi(\theta)\, dx\, d\theta \,. \tag{2.13}$$

From the definition of $\gamma^{\pi}(x)$ in (2.11), the integral in braces on the left-hand side of (2.13) is zero. As long as the set $\mathcal{A}$ receives positive probability from $f(x \mid \theta)$, the right-hand side of (2.13) is positive. Hence, we have a contradiction that shows no such relevant set $\mathcal{A}$ exists for the confidence procedure $\langle C(x), \gamma^{\pi}(x) \rangle$. A similar argument will show that, for $\langle C(x), \gamma^{\pi}(x) \rangle$, there are also no sets $\mathcal{A}$ satisfying (2.10).

Example 5. Suppose that X_i, $i = 1, 2, \ldots, n$, are independent random variables having an exponential distribution with parameter θ and we construct a confidence interval for θ. If $T = \sum_{i=1}^{n} X_i$, then $T \mid \theta$ has a gamma distribution with density

$$f(t \mid \theta) = \frac{1}{\Gamma(n)\theta^n} t^{n-1} e^{-t/\theta} \qquad t \geq 0 .$$

Observing that T/θ is a pivotal quantity, a confidence interval with coverage probability γ, is $C_E(t) = (Lt, Ut)$, where L and U are constants such that

$$\gamma = P(LT \leq \theta \leq UT \mid \theta) = P\left(\frac{1}{U} \leq \frac{T}{\theta} \leq \frac{1}{L} \,\middle|\, \theta\right) = P\left(\frac{2}{U} \leq \chi^2_{2n} \leq \frac{2}{L}\right)$$

and χ^2_{2n} is a chi-squared random variable with 2n degrees of freedom. Suppose that the prior distribution for θ is an inverted gamma density with known parameters α and β,

$$\pi(\theta \mid \alpha, \beta) \propto \frac{1}{\theta^{\alpha+1}} e^{-1/\beta\theta} \qquad \theta \geq 0$$

where $\alpha > 0$ and $\beta > 0$. The posterior distribution of θ given $T = t$ is another inverted gamma with parameters $n + \alpha$ and $(t + 1/\beta)^{-1}$, hence the posterior probability of $C_E(t)$ is

$$\gamma^{\pi}(t) = P^{\pi}(Lt \leq \theta \leq Ut \mid t) = P\left(\frac{2(t+1/\beta)}{Ut} \leq \chi^2_{2(n+\alpha)} \leq \frac{2(t+1/\beta)}{Lt}\right). \tag{2.14}$$

Note that this yields a discrepancy between a pre-data assessment of confidence and the post-data one. Indeed it easy to see that $\gamma \neq \gamma^{\pi}(t)$. Typically, posterior probabilities with respect to proper priors depend on x so they cannot be quoted as pre-data confidence. Moreover, they cannot be coverage probabilities (again, since they are functions of x). However, for $\gamma^{\pi}(x)$ of (2.14), we get a confidence procedure $\langle C(x), \gamma^{\pi}(x) \rangle$ that is free from relevant subsets.

Posterior probabilities independent of x arise if a generalised (i.e., not integrating to a finite number) prior is used, for example, if we take $\alpha = 0$ and $\beta = \infty$. Then the prior becomes $\pi(\theta) \propto \frac{1}{\theta}$, for

$\theta \geq 0$, which integrates to infinity. After applying Bayes theorem formally we obtain the posterior probability $\gamma^{\pi}(t) = P(2/U \leq \chi^2_{2n} \leq 2/L)$ which is equal to the pre-data confidence γ. For this choice of confidence there are also no relevant subsets. ||

The gamma density of Example 5 belongs to a scale family, and the prior $1/\theta$ is the standard non-informative prior for that family. In this case, equality between $\gamma^{\pi}(t)$ and γ is not surprising, as equality pre-data and post-data confidence typically occurs in invariant models (Bondar 1977), where there is often a nice correspondence between frequentist and Bayesian answers.

2.3. Kiefer's Approach

A somewhat different, but related, question was addressed by Kiefer, who detailed a methodology to merge conditional ideas with frequentist theory (Kiefer 1975, 1976, 1977, Brownie and Kiefer 1977). Kiefer was dissatisfied with the classical frequentist inference in that it could not take advantage of "lucky" or "unlucky" observations, so the report of the confidence coefficient is not data dependent. Example 4 is typical of this situation, but Kiefer's conditional evaluation of confidence was not restricted to confidence intervals. The following example describes a situation in which there are no conditional problems as in Example 4, but still a non-variable estimate of the performance of a procedure is formally correct but counterintuitive.

Example 6. Suppose that $X \sim N(\theta, 1)$ and we wish to test

$$H_0 : \theta = 1 \qquad \text{vs.} \qquad H_1 : \theta = -1 .$$

The symmetric ($\alpha = \beta$) Neyman-Pearson test rejects H_0 if $X \leq 0$ and has α level equal to 0.16 and power equal to 0.84. Thus, the pre-data confidence that we make the correct decision is 0.16, with the decision rule being the same if $x = 0.5$ or $x = 5$ (accept the null hypothesis). However, in the case $x = 5$, we should be more confident about the plausibility of H_0 than in the case $x = 0.5$. In the Neyman-Pearson theory, there is nothing that would allow us to make such a statement. ||

Example 6 is similar to Examples 3 and 4 in some aspects and different in others. The similarity with Examples 3 and 4 is that any post-data measure of "confidence" or "accuracy" should be different from the equivalent pre-data one, whether it is a measure of distance of the estimate from the

parameter, coverage probability or type I error. This is desirable since the particular values of the observed data will often alter our beliefs in the pre-data measure. Hence, one would like to use such a measure of confidence or accuracy in all three cases, and we would like such a measure to be highly dependent on the data.

However, in Examples 3 and 4, the need for a post-data measure is, in some sense, a consequence of the procedure. There exist sets on which we can condition and then the (conditional) coverage probability or risk will change for all parameter values. The sets depend on the procedure and conditioning does bring some new useful information about the coverage probability or risk. On the contrary, for Example 6, calculation of the conditional probabilities of making a type I error $P_1(\ X \leq 0 \,|\, X \in \mathcal{A})$ or of making type II error $P_{-1}(\ X > 0 \,|\, X \in \mathcal{A})$, which would be the equivalent of (2.4) or (2.5) – (2.6), is not necessarily useful, since by choosing an appropriate $\mathcal{A}$, the probability can be any value between zero and one. Thus, in Examples 3 and 4 the problem is more one of conditioning whereas in Example 6 we want a measure of achieved confidence.

2.4. Induced Partitions

Existence of conditioning sets does not necessarily mean that we are able to find them, or that they may be useful or even sensible sets. How to find them seems to be a difficult question. It is often useful to consider, instead of sets $\mathcal{A}$, more general partitions induced by a particular statistic, and search for conditioning sets which are members or unions of members of such partitions. There is a close connection between ancillarity and conditioning, and the problem at hand is no exception. Using Fisher's (1936) intuition again, ancillary statistics, that is, statistics with distribution independent of the parameter, determine the precision of an estimate without modifying its value. By thinking of the coverage probability as a measure of precision of a confidence set, it should come as no surprise that the conditional coverage probability, conditional on some partition induced by the ancillary statistic, depends on the partition.

Example 4 (continued). A version of the sufficient statistics for θ is $\left(X_{(n)}+X_{(1)},\ X_{(n)}-X_{(1)}\right)$. The distribution of the statistic $X_{(n)}-X_{(1)}$ is independent of θ, hence it is ancillary, but the statistic

$X_{(n)} + X_{(1)}$ is not sufficient for θ alone. One can compute the conditional probability

$$P_\theta\Big(\theta \in C_U(X) \,\Big|\, X_{(n)} - X_{(1)}\Big) = \min\left\{\frac{1 - \sqrt[n]{\alpha}}{1 - (X_{(n)} - X_{(1)})}, 1\right\}$$

which is a function only of $X_{(n)} - X_{(1)}$, hence it is independent of the parameter. For almost all values of $X_{(n)} - X_{(1)}$, it is different from $P_\theta\Big(\theta \in C_U(X)\Big)$, so we can immediately demonstrate (2.5) or (2.6) for all parameter values by taking $\epsilon_1 > (\sqrt[n]{\alpha} - \alpha)/(1-\alpha)$ or $\epsilon_2 < (\sqrt[n]{\alpha} - \alpha)/(1-\alpha)$, respectively. ||

Though Example 4 is a prototype example, the situation is not always so clear cut. Ancillary statistics exist only in special cases. Furthermore, an ancillary statistic might have a distribution independent of the parameter of interest, but its distribution might depend on a nuisance parameter, in which case a similar argument does not lead anywhere. For some problems ancillary statistics are not well defined, in that two statistics might individually be ancillary but jointly not so, or there exist two ancillary statistics inducing different partitions of the sample space. Then it is not obvious which partition one should use and if, by conditioning, one achieves anything. For a more thorough discussion of the above problems, as well as the relation of ancillarity with conditioning, see Basu (1964) and Kiefer (1977).

In general problems no recipe need exist, but for some other special cases there may exist a natural set to condition on. Perhaps unsurprisingly, the normal distribution with unknown mean and variance is such a case.

Example 1 (continued). For the normal distribution with both parameters unknown, using some invariance arguments (cf. Stein 1964), a crucial quantity is the ratio of sample mean $\bar{x}$ to the sample standard deviation s. For the t interval $C_t(\bar{x}, s)$, a conditioning set is

$$\mathcal{A} = \left\{x : \left|\frac{\bar{x}}{s}\right| \le K\right\} \tag{2.15}$$

for some constant K. For such a set, (2.7) is true for some positive ϵ whereas (2.8) holds by conditioning on $\mathcal{A}^c$ and setting $\epsilon = 0$, that is, if $C_t(\bar{x}, s)$ is the t confidence interval such that

$$P\Big(\mu \in C_t(\bar{X}, S) \,\Big|\, \mu, \sigma\Big) = 1 - \alpha\,, \tag{2.16}$$

then there exists positive constants K and ϵ such that

$$P\left(\mu \in C_t(\bar{X}, S) \,\middle|\, \mu, \sigma, \frac{|\bar{X}|}{S} \leq K\right) \geq 1-\alpha+\epsilon \tag{2.17}$$

and

$$P\left(\mu \in C_t(\bar{X}, S) \,\middle|\, \mu, \sigma, \frac{|\bar{X}|}{S} > K\right) \leq 1-\alpha \tag{2.18}$$

for all μ and σ (Brown 1967). ||

In the above example, the statistic $\bar{x}/s$ is a maximal invariant under the scale group. It is not obvious why such a quantity is crucial, but it is unquestionable that it contains useful information. Though no general results exist for conditioning, we would be tempted to say that it is more the invariance structure than the form of the distribution that dictates the use of the statistic.

The set $\mathcal{A}$ has an interpretation as the acceptance region of the null hypothesis $H_0: \mu = 0$. If confidence intervals are constructed only after accepting (or rejecting) the hypothesis that the mean is zero, then one implicitly conditions on the partition $\{\mathcal{A}, \mathcal{A}^c\}$. As Brown (1990) points out, the disturbing feature in a conditional evaluation is not so much the existence of some set $\mathcal{A}$ as the natural interpretation of $\mathcal{A}$ as a conditioning set. [See also the discussion between Sheffé (1977) and Olshen (1977) on Sheffé's confidence intervals, where a similar set $\mathcal{A}$ exists (Olshen 1973).] If confidence intervals were constructed in all cases the conditional performance of a procedure would be of lesser importance.

3. Constructing Measures of Conditional Confidence

The unsatisfactory conditional performance of procedures has led to several attempts to correct the problem. Most research focuses on providing a data-dependent measure of accuracy or confidence, but the exact nature of such measures vary. Bayesian statistics gives a readily available answer by simply considering the posterior probability of the parameters given the data. This is inherently post-data and does not have any pre-data interpretation. From a frequentist view, the most important attempt to construct a data-dependent measure of confidence was by Kiefer (1977), and is implicit in the conditional evaluation of procedures of Buehler (1959) and Robinson (1979a,b). The idea is to partition the sample space and report the conditional coverage probability as conditional confidence.

3.1. Conditional and Estimated Confidence

For the moment we restrict our discussion to confidence intervals, where "confidence" is translated to "probability of covering the true parameter". A pre-data confidence report γ corresponds to the best guess for the (trivially conditional) coverage probability for the coarsest possible partition, the whole sample space. Note that if the coverage probability depends on the unknown parameter θ, the report γ can be the confidence coefficient. However if (2.7) is true, we know that γ underestimates the conditional coverage probability at least by ϵ. The true value of the conditional coverage probability is typically unknown, since it often depends on the unknown parameter θ, but $\gamma+\epsilon$ is a lower bound. Hence, if $\mathrm{I}(\cdot)$ denotes an indicator function, the statistic

$$\gamma+\epsilon\,\mathrm{I}(\mathrm{x}\in\mathcal{A})=\begin{cases}\gamma+\epsilon & \text{if} \quad \mathrm{x}\in\mathcal{A}\\ \gamma & \text{if} \quad \mathrm{x}\notin\mathcal{A}\end{cases} \tag{3.1}$$

is a better estimate of the coverage probability. Reporting $\gamma+\epsilon\,\mathrm{I}(\mathrm{X}\in\mathcal{A})$ has a frequentist interpretation, since the reports are conditional probabilities. A similar construction can be used if (2.8) is true for some set.

The above example can be consider as a special case of Kiefer's conditional confidence. Kiefer extended it to different settings, such as hypothesis testing and estimation by trying to estimate quantities such as conditional probability of type I and II error or expected loss. It is worth noting that the statistic $\gamma+\epsilon\,\mathrm{I}(\mathrm{X}\in\mathcal{A})$ above can be used both as *conditional* confidence and as an *estimate* of

the confidence. The distinction is subtle, but might be clearer in the following example.

Example 7. (Kiefer 1977) Suppose that, similar to Example 1, $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$, but instead of constructing the usual t interval we construct the fixed length interval $C(\bar{x}) = [\bar{x}-c, \bar{x}-c]$. The coverage probability of $C(\bar{x})$ depends on the unknown variance and is equal to

$$P_{\mu,\sigma}\Big(\mu \in C(\bar{X})\Big) = \gamma(\sigma) = 2\,\Phi\left(\frac{\sqrt{n}\,c}{\sigma}\right) - 1\,, \tag{3.2}$$

where Φ is the standard normal cumulative distribution function. It is easy to see that $\gamma(\sigma)$ can be any number between 0 and 1. One might try to eliminate the variability in $\gamma(\sigma)$ by conditioning on a partition induced by the sample standard deviation s, with the rough idea that small values of s would indicate "lucky" observations for which the conditional coverage is larger than for large values of s. However, by conditioning on s, nothing can be achieved since the sample mean $\bar{x}$ is independent of s and the interval $C(\bar{x})$ is a function of $\bar{x}$ only.

Instead of conditioning, one can consider the coverage probability as a function of the parameter σ and try to estimate it as such. Using standard methods, it can be shown that an unbiased uniformly best estimator based on s is

$$\hat{\gamma}(s) = B_{\frac{1}{2},\frac{n-2}{2}}\left(\frac{nc^2}{s}\right), \tag{3.3}$$

where $B_{\frac{1}{2},\frac{n-2}{2}}$ is a beta density function with parameters $\frac{1}{2}$ and $\frac{n-2}{2}$. ||

Example 7 seems, however, a rather special case. In Kiefer's setup, a post-data measure of confidence is often both an estimate and a conditional confidence, though sometimes one may prove more useful than the other. It is also worth noting that in the above example the quantity to be estimated is the frequentist coverage probability, which in itself is not a post-data measure. For given data, the estimate $\hat{\gamma}(s)$ does not make a statement about the coverage of the realised confidence interval $C(\bar{x})$.

3.2. Problems with Arbitrary Partitions

The main problem with Kiefer's approach is that the partition is arbitrary, hence different

partitions would yield different answers, all of them equally valid. This becomes apparent in the testing setup (Example 6) where there does not exist a natural partition to condition on. Furthermore for an arbitrary partition, the conditional probability may depend on the unknown parameter, thus necessitating further criteria to determine valid statements (Brown 1978, Kiefer 1976).

It is worth noting that any partition does not utilise all the information that the data provide. One would expect that it is better to use more information, by asking in which member of a finer partition the sample falls. The finest partition is of course the one that consists of all possible values of the sample space. However, using the partition induced by the realised values makes a frequentist interpretation difficult, and this was considered a disadvantage by Kiefer. In the next section we will discuss in more detail the use of this finest partition.

4. Estimating Indicator Functions

The coverage probability, conditional on an arbitrary partition of the sample space, may have any value between 0 and 1. Typically, it will also be a function of the unknown parameters. Even so one may advocate a post-data estimate of confidence, however it is unlikely that it will *always* be closer to the (unknown) conditional coverage probability than the pre-data estimate. Thus to sensibly measure the closeness of the estimates we need to consider some average distance between the estimate of confidence and the conditional probability. It then seems reasonable that, instead of taking arbitrary partitions and trying to estimate conditional coverage probabilities, we should base our estimates on the finest possible partition, $\{X = x\}$. For this partition we are led immediately to the conditional probability

$$P_\theta\Big(\theta \in C(X) \,\Big|\, X = x\Big) = P_\theta\Big(\theta \in C(x)\Big) = I\Big(\theta \in C(x)\Big), \tag{4.1}$$

that is, the indicator function. The indicator function $I(\cdot)$ takes the value 0 or 1, depending on whether θ is covered by the realised confidence set C(x). Thus, $I\Big(\theta \in C(x)\Big)$ is a natural measure of post-data accuracy of C(x), measuring the confidence that we have that the realised set C(x) covers the true parameter. Since $I\Big(\theta \in C(x)\Big)$ takes the value 0 or 1, depending on θ, except in trivial cases any function of the data cannot be close to $I\Big(\theta \in C(x)\Big)$ for all values of θ. Hence, to evaluate estimators of $I\Big(\theta \in C(x)\Big)$ we use an average distance measure and, in particular, calculate a risk function.

This development which is built on ideas of Berger (1988), is in contrast to a standard frequentist pre-data assessment where the classical evaluation of the accuracy of a set estimator is through its coverage probability $P_\theta\Big(\theta \in C(X)\Big)$. This is, of course, a pre-data measure that typically depends on θ, and the usual accuracy estimate of C(X) is the confidence coefficient $\inf_\theta P\Big(\theta \in C(X)\Big)$. This pre-data measure, however, is not the best estimator of the post-data accuracy $I\Big(\theta \in C(x)\Big)$, as we will see.

4.1 Accuracy as Point Estimation

As measuring accuracy has now been equated to estimation of $I\Big(\theta \in C(x)\Big)$, we now have a problem of point estimation. Thus to assess the worth of an estimate $\gamma(x)$ of $I\Big(\theta \in C(x)\Big)$ we need to introduce a loss function. Many researchers have used squared error loss

$$L_2(\theta, \gamma) = \Big[I\Big(\theta \in C(x)\Big) - \gamma(x) \Big]^2 . \tag{4.2}$$

[Lu and Berger (1989a), George and Casella (1989), Robert and Casella (1990) Robinson (1979a,b), and Goutis and Casella (1992)]. Apart from tradition, there are a number of reasons for this choice of loss, the most interesting being that Bayes rules against (4.2) have a nice interpretation. They are the posterior probabilities of the set C(x). More precisely, for a prior $\pi(\theta)$, the Bayes rule against $L_2(\theta, \gamma)$ is given by

$$\gamma^{\pi}(x) = P\Big(\theta \in C(x) \,\Big|\, x\Big) = \int_{C(x)} \pi(\theta \,|\, x)\, d\theta \tag{4.3}$$

where $\pi(\theta \,|\, x)$ is the posterior distribution of θ as in Example 5. Loss functions having this property are called proper (Lindley 1985). The idea of using a quadratic loss for estimating an indicator function is quite old (Brier 1950), and proper loss functions have been studied by Savage (1971), who provides further references, and more recently by Lindley (1982, 1985) and Schervish (1989). From a Bayesian point of view, their use is advocated for assessing subjective probabilities since, if the loss is proper, it turns out that the best strategy for a probability assessor is to quote the true personal probability.

The quadratic loss function is by no means the only proper loss function. However, the results of Hwang and Pemantle (1990), though applicable in the estimation of a fixed indicator function, suggest that the quadratic loss plays a key role among all proper loss functions.

Once the loss is specified, an estimator $\gamma(X)$ can be evaluated in terms of risk

$$R(\theta, \gamma) = E_{\theta}\Big[I\Big(\theta \in C(X)\Big) - \gamma(X) \Big]^2 . \tag{4.4}$$

An estimator $\gamma_1(X)$ is better than another estimator $\gamma_2(X)$ if it has smaller risk for all parameter values. Traditional decision theoretic criteria such as admissibility, minimaxity etc. can be used, though the technical difficulties are appreciable since $\gamma(X)$ estimates a random function. General decision theoretic results are not directly applicable, but one can easily show that proper or, sometimes, generalised Bayes probabilities cannot be dominated in risk by any other estimator.

4.2. Examples

A number of researchers have constructed estimates of $I\Big(\theta \in C(x)\Big)$, and then demonstrated the

superiority of their estimates over the usual pre-data assessments. Construction of such estimators is sometimes technically involved, so we will only describe some results with minimal details.

Example 8 (multivariate normal). Suppose $\mathbf{X} = \mathbf{x}$ is an observation from a p-variate normal distribution with mean θ and identity covariance matrix, $\mathbf{X} \sim \mathrm{N}_p(\theta, \mathbf{I})$. The classical confidence interval is

$$C_0(x) = \{\theta : \|\theta - x\| \le c\}, \tag{4.5}$$

where c is chosen to satisfy $P(\chi^2_p \le c^2) = 1-\alpha$, where $1-\alpha$ is a specified level and χ^2_p denotes a chi-squared random variable with p degrees of freedom. For $p \ge 4$, this interval is dominated by

$$C_+(x) = \left\{\theta : \left\|\theta - \delta^+(x)\right\| \le c\right\} \tag{4.6}$$

where

$$\delta^+(x) = \left(1 - \frac{p-2}{\|x\|^2}\right)^+ x, \tag{4.7}$$

the positive-part Stein estimator (Hwang and Casella 1982). This domination is of the form

$$P_\theta\Big(\theta \in C_+(X)\Big) > P_\theta\Big(\theta \in C_0(X)\Big) \tag{4.8}$$

for all θ, hence is a pre-data domination. In terms of post-data inference, however, classical frequency gives us the same post-data inference for $C_+(x)$ and $C_0(x)$. Since $\inf_\theta P\Big(\theta \in C_+(X)\Big) = 1-\alpha$, both sets have $1-\alpha$ as a frequentist confidence coefficient. ||

Does there exist an estimate of confidence $\gamma(x)$ for $C_+(x)$ that dominates $1-\alpha$ in the sense of (4.4)? The question was addressed by Lu and Berger (1989a) who showed that

$$\gamma_{LB}(x) = 1 - \alpha + \frac{a}{b + \|x\|^2}, \tag{4.9}$$

where a and b are positive constants with $a \le b\alpha$, dominates $1-\alpha$ as an assessment of accuracy of $C_+(x)$ under squared error loss. Subsequently George and Casella (1989) constructed the accuracy estimate

$$\gamma_{EB}(x) = P\left(\chi^2_p \le \frac{c^2}{u_{a,b}(\|X\|^2)}\right), \tag{4.10}$$

where, for some constants $a > 0$ and $b \in [0, 1)$,

$$u_{a,b}(r) = \max\left\{\left(1 - \frac{a}{r}\right), b\right\}, \tag{4.11}$$

as a post-data accuracy measure of $I\Big(\theta \in C_+(x)\Big)$. This estimate was derived using empirical Bayes methods and was also shown to dominate $1-\alpha$.

The estimator $\gamma_{EB}(x)$ was constructed as an empirical Bayes version of a Bayesian posterior probability. As previously mentioned, such an estimator is expected to be a good estimate of the indicator function. Interestingly, we can think of $\gamma_{LB}(x)$ as a Taylor series approximation to this estimate. It also turns out that the expected values of these estimates somewhat resemble the coverage probability $P_\theta\Big(\theta \in C_+(X)\Big)$. Thus we have not only improved our post-data estimate, but also have a reasonable pre-data estimate. Figure 1 illustrates this behaviour.

Example 8 (continued). Since $C_+(x)$ has a nonconstant coverage probability, it was expected that $1-\alpha$ could be improved upon as a post-data accuracy estimate. For the set $C_0(x)$ however, it is not clear that we can improve on $1-\alpha$ as a post-data accuracy estimate. Surprisingly, we can, as demonstrated by Robert and Casella (1990). Using both empirical Bayes arguments and Taylor series approximations they derived the accuracy estimate

$$\gamma^*(x) = 1-\alpha+\frac{a}{\|x\|^2}, \tag{4.12}$$

where a is a nonnegative constant. They demonstrated that, for $p \geq 5$, $\gamma^*(x)$ is a better post-data accuracy measure of C_0 than $1-\alpha$. ||

An interesting feature of the accuracy estimators of Example 8 is that there is no sample space partition on which they are based. Indeed, it is not clear if any such partition exists, but happily, the methodology of estimation of indicator functions is not dependent on the existence of partition. When one does exist, however, it may be possible to take advantage of it.

Example 1 (continued). For the usual $1-\alpha$ Student's t interval $C_t(\bar{x}, s^2)$, Brown (1967) demonstrated the existence of positive constants K and ϵ such that

$$P\left(\mu \in C_t(\bar{X}, S) \,\middle|\, \mu, \sigma^2, \frac{|\bar{X}|}{S} \leq K\right) > 1-\alpha+\epsilon \tag{4.13}$$

for all μ and σ^2. This determines the partition based on the set $\mathcal{A} = \{x : |\bar{x}|/s \leq K\}$ and, as in (3.1), this information can be used to construct improved estimates of post-data accuracy. Goutis and

Casella (1992) addressed the problem, and using this conditioning set, combined with limiting arguments, constructed the estimate

$$\gamma(\bar{x}/s) = \begin{cases} \dfrac{1-\alpha}{P(|t_{n-1}| < (|\bar{x}|/s)\sqrt{n-1})} & \text{if} \quad |\bar{x}|/s > c^* \\ \dfrac{1-\alpha}{P(|t_{n-1}| < c^*\sqrt{n-1})} & \text{if} \quad |\bar{x}|/s \le c^* \end{cases}, \tag{4.14}$$

where t_{n-1} is a random variable having a t distribution with $n-1$ degrees of freedom and c^* a constant. They showed that $\gamma(\bar{x}/s)$, which is uniformly larger than $1-\alpha$, is a better measure of the post-data accuracy of $C_t(\bar{x}, s)$ for $n > 2$. ||

There is a subtle difference between the partition of Example 1 and the partitions used by Kiefer (1977). In the former case there is really no arbitrariness in the partition, as it is induced by the procedure itself. In contrast, Kiefer's partitions were induced by the experimenter. Although one might argue that such arbitrariness is desirable, as it allows one to obtain precise inferences, it ignores conditional properties of the set estimator itself. These conditional properties, if they are evident, should always be the basis of constructing post-data measures.

4.3. Other considerations

An additional requirement proposed sometimes (Berger 1985a, Lu and Berger 1989a, Hwang and Brown 1991) is that the estimator $\gamma(x)$ is *frequentist valid*, that is,

$$E_\theta\, \gamma(X) \le P_\theta\Big(\theta \in C(X)\Big), \qquad \text{for all } \theta\,. \tag{4.15}$$

The frequentist validity criterion expresses a need to be conservative and is justifiable if we consider the coverage/noncoverage of a given confidence set as a $0-1$ loss. Then one would not want to report an over-optimistic estimate of the obtained loss. A closely related idea is the guaranteed confidence (Brown 1978).

It is worth noting the close relation between estimation of an indicator function using quadratic loss and the report of conditional confidence as developed in Section 3. The following result (Robinson 1979a) shows that if (2.7) is true for some set $\mathcal{A}$, the estimate $\gamma + \epsilon\, I(x \in \mathcal{A})$, which was derived in an intuitive way (Section 3.1), has a better risk than γ. The difference in risks is given by

$$
\begin{aligned}
E_\theta\Big[\, I\big(\theta \in C(X)\big) - \gamma \,\Big]^2 - E_\theta\Big[I\big(\theta \in C(X)\big) - \gamma - \epsilon\, I(X \in \mathcal{A}) \,\Big]^2 & \\
&\geq 2\epsilon\, E_\theta\Big[I\big(\theta \in C(X),\, X \in \mathcal{A}\big) - (\gamma + \epsilon)\, I(X \in A)\Big] \\
&= 2\epsilon\, P_\theta(X \in \mathcal{A})\Big\{ P_\theta\big(\theta \in C(X) \,|\, X \in \mathcal{A}\big) - (\gamma + \epsilon)\Big\} \\
&> 0\,, \qquad (4.16)
\end{aligned}
$$

showing that $\gamma + \epsilon\, I(x \in \mathcal{A})$ dominates γ. The estimate $\gamma + \epsilon\, I(x \in \mathcal{A})$ was derived to estimate the coverage probability conditioning on the partition $\{\mathcal{A}, \mathcal{A}^c\}$. Indeed if $x \in \mathcal{A}$, $\gamma + \epsilon$ is always closer than γ to the conditional probability, since it is a lower bound. The existence of a partition such that we can find a useful non-trivial lower (or upper) bound is not a general phenomenon. It is essentially equivalent with the existence of relevant sets.

5. Other Accuracy Measures

Estimation of the indicator of coverage can be considered as a special case of the more general subject of loss estimation. In this section we look at the general case, as well as some other special cases. In particular, we discuss the formulation of accuracy estimation in testing.

5.1. Loss Estimation

A somewhat similar, but not identical problem is estimating the loss of a point estimator. Consider $\delta(x)$ a point estimator of a parameter θ and suppose that we evaluate $\delta(x)$ according to some loss $L\Big(\theta, \delta(x)\Big)$, say a measure of the distance between $\delta(x)$ and the true value of θ. Then a pre-data measurement of the performance of $\delta(X)$ is the risk

$$R(\theta, \delta) = E_\theta\Big[L\Big(\theta, \delta(X)\Big)\Big], \tag{5.1}$$

which in general is a function of the unknown parameter θ. A conservative pre-data report of the goodness of an estimator could be $\sup_\theta R(\theta, \delta)$, which measures how far from θ we expect the estimator to be, in the worst case. Once the data are obtained, interest lies on the accuracy of the obtained estimate $\delta(x)$, hence the risk might not be the appropriate quantity. Furthermore, Example 3 shows that the behaviour of the conditional risk may be different from that of the unconditional one. It can be argued in a similar way to the previous section that a better quantity to estimate is the observed risk, i.e., the risk conditional in the observed data. Of course this is nothing but the loss, hence we want an estimate $\hat{L}(x)$ of the attained loss $L\Big(\theta, \delta(x)\Big)$. We can now measure the performance of $\hat{L}(x)$ by a measure of distance of $\hat{L}(x)$ from the true loss. For example, we can use

$$L^*(\theta, \delta) = \Big[\hat{L}(x) - L\Big(\theta, \delta(x)\Big)\Big]^2, \tag{5.2}$$

and compare different loss estimates using the risk

$$R^*(\theta, \delta) = E_\theta\Big[\hat{L}(x) - L\Big(\theta, \delta(x)\Big)\Big]^2. \tag{5.3}$$

This approach was taken by Sandved (1968) but, somehow curiously, it was forgotten until Kiefer (1977) tried to embed it in the general theory of estimated confidence. Later it was considered by Berger (1985a,b), Lu and Berger (1989b) and Johnstone (1988). However it seems that it has not

received the attention that it probably deserves.

An interesting way of combining the estimation loss of $\delta(x)$ with the estimated precision can be found in Rukhin (1988a,b,c). Rukhin's decision-precision approach considers a procedure good if the $\delta(x)$ is close to the parameter θ that it estimates in some distance sense, measured by a loss $W(\theta, \delta)$ and if another component γ of the loss measures the precision of $\delta(x)$, that is, it estimates $W(\theta, \delta)$. A decision-precision loss is

$$L(\theta, \delta, \gamma) = W(\theta, \delta)\gamma^{-1/2} + \gamma^{1/2} , \tag{5.4}$$

which turns out to have several attractive features (see Rukhin 1988a for details).

5.2. Estimation of Accuracy in Testing.

The problem of hypothesis testing can be viewed in the same spirit. For a test of

$$H_0 : \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \theta \notin \Theta_0 , \tag{5.5}$$

based on observing $X = x$, where $X \sim f(x \mid \theta)$, estimation of the quantity $I(\theta \in \Theta_0)$ can assess the accuracy of the testing procedure. A pre-data assessment of a test of (5.5) would involve calculation of Type I and Type II errors. If a rejection rule for (5.5) is "Reject H_0 if $x \in R$", then the type I error is

$$P(X \in R \mid \theta \in \Theta_0) = P_\theta(X \in R)\ I(\theta \in \Theta_0) . \tag{5.6}$$

This is a pre-data quantity, a parameter that could be estimated. However, once $X = x$ is observed we are interested in the post-data version of (5.6), which is

$$P_\theta(x \in R)\ I(\theta \in \Theta_0) = I(x \in R)\ I(\theta \in \Theta_0) . \tag{5.7}$$

Since both x and R are known, the value of $I(x \in R)$ is known. Thus we are left with the unknown parameter $I(\theta \in \Theta_0)$, which measures our post-data accuracy.

Another argument, although somewhat less compelling, also leads to the conclusion that estimation of $I(\theta \in \Theta_0)$ is reasonable. An advantage of this latter argument is that it shows us ways to measure the accuracy of estimation of $I(\theta \in \Theta_0)$. The argument is quite simple: classical hypothesis testing is a special case of estimating $I(\theta \in \Theta_0)$.

If we start with observing $X = x$ and we want to estimate $I(\theta \in \Theta_0)$ by a function $\phi(x)$ under the loss

$$L(\theta, \phi) = |I(\theta \in \Theta_0) - \phi(x)|, \tag{5.8}$$

the Bayes rule for a prior $\pi(\theta)$ is

$$\phi^{\pi}(x) = \begin{cases} 1 & \text{if} \quad P(\theta \in \Theta_0 | x) \geq 1/2 \\ 0 & \text{if} \quad P(\theta \in \Theta_0 | x) < 1/2 \end{cases}. \tag{5.9}$$

The estimate $\phi^{\pi}(x)$ is actually a Neyman-Pearson type rejection rule. [Note our departure from the usual notation for Neyman-Pearson critical function which, for nonrandomised test, is $1-\phi(x)$.] Furthermore, Type I and Type II errors of a $0-1$ rule $\phi(x)$ are the risks with respect to the loss (5.8) for $\theta \in \Theta_0$ and $\theta \notin \Theta_0$. Thus classical hypothesis testing can be regarded as estimation using the loss function (5.8).

However, the form of the loss (5.8) and the general fact that optimal procedures must be Bayes procedures of some kind, restrict the rules $\phi(x)$ (whether Bayes or Neyman-Pearson rules) to be $0-1$ functions. As we saw in Example 6, a disadvantage is that for all x in the rejection region, the same $\phi(x)$ is reported. Hence the form of the loss restricts the rules to be the pre-data probabilities of errors. Few people may share DeFinetti's (1974) opinion that "accept or reject is the unhappy formulation which [he considers] as the principal cause of the fogginess widespread all over the field of statistical inference and general reasoning", but the unhappy formulation might explain the popularity among practitioners of data-dependent measures such as p-values, which, after all, have not had as formal a development.

The form of the loss function immediately suggests examining the class of loss functions

$$L_k(\theta, \phi) = |I(\theta \in \Theta_0) - \phi(x)|^k, \qquad k = 1, 2, \cdots. \tag{5.10}$$

This was done, to a certain extent by Hwang *et al.* (1992) and Goutis *et al.* (1993), with the conclusion that $L_2(\theta, \phi)$ was one of the most interesting losses. Similar to the set estimation problem, Bayes rules against $L_2(\theta, \phi)$ are posterior probabilities, i.e., instead of the $0-1$ rule given by (5.9), one obtains $\phi^{\pi}(x) = P(\theta \in \Theta_0 | x)$. Other proper loss functions would also yield posterior probabilities as optimal rules.

Having an estimator of $I(\theta \in \Theta_0)$ that has an interpretation as a probability is quite nice. This allows the experimenter to make a post-data probabilistic assessment of a validity of H_0. This seems

to be what is done with a p-value (and is in many situations) but now we have a legitimate formal framework in which to make inference.

It is worth noting that treating the hypothesis testing problem as an estimation of $I(\theta \in \Theta_0)$ allows us to report highly data dependent measures of evidence for or against H_0, but it does not seem at first to be post-data estimation of the accuracy of procedures in the spirit of estimation of $I\Big(\theta \in C(x)\Big)$ for confidence sets or $L\Big(\theta, \delta(x)\Big)$ for point estimates. The approach seems to have more in common with estimating a particular function of the parameter θ, than with estimated or conditional confidence.

Treating the problem as one of estimating confidence of the procedure, Kiefer (1977), as a part of his general theory, constructed measures of post-data confidence. Perhaps it easy to understand his constructions by using Example 6.

Example 6 (continued). One might try to improve upon the unconditional Neyman-Pearson procedure by considering a partition of the sample space finer than $\{(-\infty, 0], (0, \infty)\}$ which corresponds to the rejection and acceptance regions. Suppose that the decision rule remains the same but we consider that $|x| > 1$ indicates strong evidence against the rejected hypothesis whereas $|x| \leq 1$ indicates weak evidence. This induces the partition $\{(-\infty, -1), [-1, 0], (0, 1], (1, \infty)\}$. Then, depending on whether $|x| > 1$ or $|x| \leq 1$ we report the conditional confidence coefficient

$$P_{\pm 1}(\text{making the correct decision} \,\big|\, |X| > 1) \tag{5.11}$$

or

$$P_{\pm 1}(\text{making the correct decision} \,\big|\, |X| \leq 1)\ . \tag{5.12}$$

Hence, if $x = 5$ the decision rule is to accept H_0, but the conditional confidence is $P_1(X > 0 \,\big|\, |X| > 1) = 0.96$ whereas $x = 0.5$ yields a confidence equal to $P_1(X > 0 \,\big|\, |X| \leq 1) = 0.71$. ||

It is worth noting that, in the above example, one could have considered "strong evidence against the rejected hypothesis" the event $|x| > 2$, instead of $|x| > 1$. In that case the conditional confidence would be different. Instead of computing the conditional confidence, one might try to estimate the probability of making the correct decision, as a function of the parameter. Consider the following

variation of Example 6.

Example 9. Suppose that $X \sim N(\theta, 1)$ and we wish to test

$$H_0 : \theta \leq 0 \qquad \text{vs.} \qquad H_1 : \theta > 0 \,. \tag{5.13}$$

The symmetric Neyman-Pearson test rejects H_0 if $x \leq 0$ but both probabilities of type I and type II errors approach 1/2 as $|\theta| \to 0$. The true probabilities are functions of the parameters and are equal to $\Phi(|\theta|)$, where Φ is the standard normal cumulative distribution function. Large values of $|x|$ indicate that $|\theta|$ is far from 0, so, intuitively, we should be more certain that we made the correct decision (accept or reject) if $|x|$ is large. One way to quantify this intuition is by estimating the probability of error. The maximum likelihood estimate of $\Phi(|\theta|)$ is $\Phi(|x|)$, which could be reported as a measure of the estimated confidence. ||

It is interesting that data-dependent (conditional) estimation of power is mentioned in Lehmann (1986, p.151) in the treatment of tests for multiparameter exponential families, but not advocated because it is not clear on which set one should condition and, furthermore, the estimator becomes available only after the observations are taken, hence it cannot be used to plan an experiment.

Estimated confidence in testing is not concerned, however, with the realised data and rejection or acceptance depending on the data. In Example 9, the concern is to estimate the risk of the test, as opposed with the attained loss. We can however develop the problem of estimation the loss as follows (Casella and Goutis 1992). For the testing problem of (5.5), we consider a testing rule of the form: accept H_0 if $x \in A$, where A is some region in the sample space (the acceptance region), that is, we take $\phi(x) = I(x \in A)$. The loss (5.8) incurred by this procedure can be written

$$L\Big(\theta, \phi(x)\Big) = \begin{cases} 1 & \text{if } \theta \in \Theta_0,\ x \notin A \text{ or } \theta \notin \Theta_0,\ x \in A \\ 0 & \text{if } \theta \in \Theta_0,\ x \in A \text{ or } \theta \notin \Theta_0,\ x \notin A \end{cases} . \tag{5.14}$$

To estimate the loss (or accuracy) of the acceptance region, we use estimators p(x) that perform well against the loss

$$\begin{aligned} L^*\Big(\theta, \phi(x)\Big) &= \Big[L\Big(\theta, \phi(x)\Big) - p(x)\Big]^2 \\ &= \Big[\big|I(\theta \in \Theta_0) - I(x \in A)\big| - p(x)\Big]^2 . \end{aligned} \tag{5.15}$$

Given a prior π, the Bayes rule against the $L^*\Big(\theta, \phi(x)\Big)$ (for a given acceptance region A) is

$$p^{\pi}(x) = P(\theta \in \Theta_0 \,|\, x)\; I(x \notin A) + P(\theta \notin \Theta_0 \,|\, x)\; I(x \in A)\;. \tag{5.16}$$

Hence, if the hypothesis test rejects H_0 [so that $I(x \in A) = 0$] the estimated loss of the test is the posterior probability of Θ_0, whereas if H_0 is accepted $\Big(I(x \in A) = 1\Big)$ then $p^{\pi}(x)$ equals the posterior probability of Θ_0^c, and gives the evidence against H_0. Of course the estimator $p(x)$ need not necessarily be a Bayes estimator, but (5.16) suggests that, if $\delta(x)$ is any estimate of $I(\theta \in \Theta_0)$, a reasonable estimate of $L\Big(\theta, \phi(x)\Big)$ is

$$p(x) = \delta(x)\; I(x \in A^c) + \Big(1 - \delta(x)\Big)\, I(x \in A)\;. \tag{5.17}$$

Then it turns out that

$$\begin{aligned} L^*\Big(\theta, \phi(x)\Big) &= \Big(\,|\, I(\theta \in \Theta_0) - I(x \in A)\,|\, - p(x)\Big)^2 \\ &= \Big(I(\theta \in \Theta_0) - \delta(x)\Big)^2 \\ &= L_2(\theta, \delta)\;, \end{aligned} \tag{5.18}$$

and, hence, the loss functions are equivalent and estimators of accuracy of an acceptance region A can be constructed from estimators of $I(\theta \in \Theta_0)$, and the risks will be the same.

6. Discussion

Throughout this paper we have been concerned with methods for frequentist post-data inference. This concern, we hope, has not overshadowed the importance of the pre-data/post-data distinction. In any experiment both pre-data inferences and post-data inferences are important, and each can be made within either frequentist or Bayesian paradigms, which perhaps shows that the frequentist/Bayesian distinction is not as fundamental as the pre-data/post-data distinction. The result of a statistical experiment is an inference. By focusing on inference, and whether the inference is pre-data or post-data, we are concentrating on the most important consequence of the experiment.

Post-data inference has traditionally been the inference of Bayesian statistics. In fact there have been criticisms of frequentist post-data inference that have essentially said "why bother?", as the Bayesians have methods. Again, the reason to bother is because post-data inference is an inference that can be used by any statistical school.

The methodologies outlined in Sections 4 and 5 provide a reasonably comprehensive approach for performing frequentist post-data inference. By reducing the problem to, essentially, one of point estimation, standard frequency methods of evaluation can be used. The one difficulty which we believe has been satisfactorily addressed is to identify an object of estimation (parameter) that measures post-data accuracy. The indicator of coverage, and variants in Section 5, seem to measure accuracy quite well. Moreover, there are several situations when a pre-data conservative report of confidence coefficient is undesirable for other reasons. Several improved confidence sets have been constructed, such as in Example 8, in which the improvement over the standard sets yields a higher coverage probability with the same or smaller volume. Typically, the confidence coefficient is equal to that of the standard set, but reporting the confidence coefficient does not make the attained gains in coverage probability tangible.

A final observation we would like to make is that a successful statistical procedure must be successful in both pre-data and post-data uses. Such a procedure must, therefore, fare well when evaluated conditionally or unconditionally. This then translates into a procedure performing reasonably under both Bayesian and frequentist evaluations. Ignoring either one of these evaluations will result in a less-than-optimal answer.

Appendix

Proof of inequality (2.4). The difference of the risks is

$$\begin{aligned}
&\mathrm{E}_\theta\Big(\|\delta(\mathrm{X})-\theta\|^2\,\Big|\,\|\mathrm{X}\|^2\le\epsilon\Big)-\mathrm{E}_\theta\Big(\|\mathrm{X}-\theta\|^2\,\Big|\,\|\mathrm{X}\|^2\le\epsilon\Big)\\
&\qquad=(\mathrm{p}-2)^2\,\mathrm{E}_\theta\Big(\frac{1}{\|\mathrm{X}\|^2}\,\Big|\,\|\mathrm{X}\|^2\le\epsilon\Big)-2\,(\mathrm{p}-2)\,\mathrm{E}_\theta\Big(\frac{\mathrm{X}'(\mathrm{X}-\theta)}{\|\mathrm{X}\|^2}\,\Big|\,\|\mathrm{X}\|^2\le\epsilon\Big)\\
&\qquad=(\mathrm{p}-2)\Big[(\mathrm{p}-2)\,\mathrm{E}_\theta\Big(\frac{1}{\|\mathrm{X}\|^2}\,\Big|\,\|\mathrm{X}\|^2\le\epsilon\Big)-2+2\,\mathrm{E}_\theta\Big(\frac{\mathrm{X}'\theta}{\|\mathrm{X}\|^2}\,\Big|\,\|\mathrm{X}\|^2\le\epsilon\Big)\Big]\\
&\qquad\ge(\mathrm{p}-2)\Big[\frac{\mathrm{p}-2}{\epsilon}-2+2\,\mathrm{E}_\theta\Big(\frac{\mathrm{X}'\theta}{\|\mathrm{X}\|^2}\,\Big|\,\|\mathrm{X}\|^2\le\epsilon\Big)\Big]\,. \qquad \text{(A.1)}
\end{aligned}$$

For $\epsilon<(\mathrm{p}-2)/2$ it suffices to show

$$\mathrm{E}_\theta\Big(\frac{\mathrm{X}'\theta}{\|\mathrm{X}\|^2}\,\Big|\,\|\mathrm{X}\|^2\le\epsilon\Big)\ge 0\,. \qquad \text{(A.2)}$$

Since the LHS of (A.2) is rotation invariant we can take without loss of generality $\theta=(\theta_1,\,0,\,0,\ldots,\,0)$ where $\theta_1\ge 0$. Observing that $\mathrm{X}'\theta/\|\mathrm{X}\|^2$ is an odd function of X_1, and $\mathrm{P}_\theta\Big(\mathrm{X}_1\ge 0\,\Big|\,\|\mathrm{X}\|^2\le\epsilon\Big)\ge 1/2$, the expectation in (A.2) is positive, establishing (2.4).

REFERENCES

Basu, D. (1964) Recovery of ancillary information. *Contributions to Statistics*, 7-20. Pergamon Press, Oxford. Also in *Statistical Information and Likelihood. A Collection of Critical Essays by Dr. D. Basu*, 1988, J. K. Ghosh (ed.). Lecture Notes in Statistics, Springer-Verlag, New York.

Berger, J. O. (1985a) The frequentist viewpoint and conditioning. *Proc. Berkeley Conference in Honor of Jersy Neyman and Jack Kiefer* (L. M. LeCam and R. A. Olshen, eds.) **1** 15-44 Wadsworth, Monterey, California.

Berger, J. O. (1985b) In defence of the likelihood principle: axiomatics and coherency. In *Bayesian Statistics* 2 (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 33-65 North Holland, Amsterdam.

Berger, J. O. (1988) An alternative: The estimated confidence approach. In *Statistical Decision Theory and Related Topics IV, Vol. 1* (S. S. Gupta and J. O. Berger, eds.) 85-90. Springer Verlag, New York.

Berger, J. O. and Wolpert, R. L. (1988) *The Likelihood Principle, 2nd Edition.* IMS monograph series. Institute of Mathematical Statistics, Hayward, California.

Bondar, J. V. (1977) A conditional confidence principle. *Ann. Statist.* **5** 881-891.

Brier, G. W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78** 1-3.

Brown, L. D. (1967) The conditional level of Student's t test. *Ann. Math. Statist.* **38** 1068-1071.

Brown, L. D. (1978) A contribution to Kiefer's theory of conditional confidence procedures. *Ann. Statist.* **6** 59-71.

Brown, L. D. (1990) Comment on "Developments in decision theoretic variance estimation" by J. M. Maatta and G. Casella. *Statist. Sci.* **5** 90-120.

Brownie, C. and Kiefer, J. (1977) The ideas of conditional confidence in the simplest setting. *Commun. Statist. Theor. Meth.* **A6** 691-751.

Buehler, R. J. (1959) Some validity criteria for statistical inference. *Ann. Math. Statist.* **30** 1068-107.

Buehler, R. J. and Feddersen, A. P. (1963) Note on a conditional property of Student's t. *Ann. Math. Statist.* **34** 1098-1100.

Casella, G. (1992) Conditional inference from confidence sets. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu* (M. Ghosh and P. K. Pathak eds.) 1-12. IMS monograph series. Institute of Mathematical Statistics, Hayward, California.

Casella, G. and Goutis, C. (1992) Relationships between post-data accuracy measures. Biometrics Unit Technical Report BU-1047-M, Cornell University, Ithaca, New York.

DeFinetti, B. (1974) Bayesianism: its unifying role in the foundations and applications of statistics. *Int. Statist. Rev.* **42** 117-130.

Fisher, R. A. (1936) Uncertain inference. *Proceedings of the American Academy of Arts and Sciences* **71** 245-258.

Fisher, R. A. (1959) *Statistical Methods and Scientific Inference, 2nd Edition.* Hafner, New York.

George, E. I. and Casella, G. (1989) Empirical Bayes confidence estimation. Biometrics Unit Technical Report BU-1062-M, Cornell University, Ithaca, New York. To appear in *Statistica Sinica.*

Goutis, C. and Casella, G. (1992) Increasing the confidence in Student's t-interval. *Ann. Statist.* **20** 1501-1513.

Goutis, C., Casella, G. and Wells, M. T. (1993) Assessing evidence in multiple hypothesis. Biometrics Unit Technical Report BU-1084-M, Cornell University, Ithaca, New York.

Hwang, J. T. and Brown, L. D. (1991) Estimated confidence under the validity constraint. *Ann. Statist.* **19** 1964-1977.

Hwang, J. T. and Casella, G. (1982) Minimax confidence sets for the mean of a multivariate normal distribution. *Ann. Statist.* **10** 868-881.

Hwang, J. T., Casella, G., Robert, C., Wells, M. T. and Farrell, R. H. (1992) Estimation of accuracy in testing. *Ann. Statist.* **20** 490-509.

Hwang, J. T. and Pemantle, R. L. (1990) Evaluation of estimators of statistical significance under a class if proper loss functions. Technical Report, Statistics Center, Cornell University, Ithaca, New York.

James, W. and Stein, C. (1961) Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 361-380.

Johnstone, I. (1988) On inadmissibility of some unbiased estimates of loss. In *Statistical Decision Theory and Related Topics* IV, Vol. 1 (S. S. Gupta and J. O. Berger, eds.) 361-379. Springer Verlag, New York.

Kiefer, J. (1975) Conditional confidence and estimated confidence in multidecision problems (with applications to selection and ranking). In *Multivariate Analysis* IV (P. R. Krishnaiah, ed.). North Holland, Amsterdam.

Kiefer, J. (1976) Admissibility of conditional confidence procedures. *Ann. Statist.* **4** 836-865.

Kiefer, J. (1977) Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72** 789-827.

Lehmann, E. L. (1986) *Testing Statistical Hypotheses, 2nd Edition.* John Wiley, New York.

Lindley, D. V. (1982) Scoring rules and the inevitability of probability (with discussion) *Int. Statist. Rev.* **50** 1-26.

Lindley, D. V. (1985) *Making Decisions, 2nd Edition.* John Wiley, London.

Lu, K. L. and Berger, J. O. (1989a) Estimated confidence procedures for multivariate normal means. *J. Statist. Plan. Inf.* **23** 1-19.

Lu, K. L. and Berger, J. O. (1989b) Estimation of normal means: frequentist estimation of loss. *Ann. Statist.* **17** 890-906.

Maatta, J. M. and Casella, G. (1987) Conditional properties of interval estimators of a normal variance. *Ann. Statist.* **15** 1372-1388.

Olshen, R. A. (1973) The conditional level of F-test. *J. Amer. Statist. Assoc.* **72** 789-827.

Olshen, R. A. (1977) A note on the reformulation of the S-method of multiple comparison. *J. Amer. Statist. Assoc.* **72** 144-146.

Pierce, D. A. (1973) On some difficulties in a frequency theory of inference. *Ann. Statist.* **1** 241-250.

Robert, C. and Casella, G. (1990) Improved confidence estimators for the usual multivariate normal confidence set. Biometrics Unit Technical Report BU-1041-M, Cornell University, Ithaca, New York. To appear in *Statistical Decision Theory and Related Topics* V (S. S. Gupta and J. O. Berger, eds.).

Robinson, G. K. (1975) Some counterexamples to the theory of confidence intervals. *Biometrika* **62** 155-161.

Robinson, G. K. (1979a) Conditional properties of statistical procedures. *Ann. Statist.* **7** 742-755.

Robinson, G. K. (1979b) Conditional properties of statistical procedures for location and scale families. *Ann. Statist.* **7** 756-771.

Rukhin, A. L. (1988a) Estimated loss and admissible loss estimators. In *Statistical Decision Theory and Related Topics* IV, Vol. 1 (S. S. Gupta and J. O. Berger, eds.). 409-418. Springer Verlag, New York.

Rukhin, A. L. (1988b) Estimating the loss of estimators of a binomial parameter. *Biometrika* **75** 153-155.

Rukhin, A. L. (1988c) Loss functions for loss estimation. *Ann. Statist.* **16** 1262-1269.

Savage, L. J. (1971) Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783-801.

Sandved, E. (1968) Ancillary statistics and estimation of the loss in estimation problems. *Ann. Math. Statist.* **39** 1756-1758.

Schervish, M. J. (1989) A general method for comparing probability assessors. *Ann. Statist.* **17** 1856-1879.

Sheffé, H. (1977) A note on the reformulation of the S-method of multiple comparison. *J. Amer. Statist. Assoc.* **72** 143, 146.

Stein, C. (1964) Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.* **16** 155-160.

Wallace, D. L. (1959) Conditional confidence level properties. *Ann. Math. Statist.* **30** 864-876.

Figure 1: Coverage probability of $C_+(X)$ (dotted lines) and expected values of confidence estimators $\gamma_{EB}(X)$ (solid lines) and $\gamma_{LB}(X)$ (dashed lines) for $1-\alpha = 0.9$. The constants used for $\gamma_{EB}(X)$ are $a = (p-2)^2/p$ and $b = 2c^2\left(c^2 + (p-2) + c\sqrt{c^2 + 4(p-2)}\right)^{-1}$, and for $\gamma_{LB}(X)$ they are $a = (p-2)^2 c^p e^{-c^2/2} 2^{-p/2} p^{-1}[\Gamma(p/2)]^{-1}$ and $b = a/\alpha$. Both $\gamma_{EB}(X)$ and $\gamma_{LB}(X)$ are truncated at $\gamma_{max} = \max_{\theta} P_\theta\left(\theta \in C_+(X)\right)$.